International Journal of Bio-Technology
and Research (IJBTR)
ISSN(P): 2249-6858; ISSN(E):2249-694X
Vol. 3, Issue 5, Dec 2013, 1-10

STOCHASTIC CONTEXT-FREE GRAMMARS METHOD AIDED CALCULATION OF THE LOCAL FOLDING POTENTIAL OF TARGET RNA

SHAHIRA M. HABASHY

Faculty of Engineering, Helwan University, Cairo, Egypt

ABSTRACT

RNA structure is an important field of study. Prediction structure can overcome many of the issues with physical structure determination. For several decades, free energy minimization has been the most popular method for prediction from a single sequence. It is based on a set of empirical free energy change parameters derived from experiments using a nearest-neighbor model. Accurate prediction of RNA secondary structure from the base sequence is an unsolved computational challenge. The accuracy of predictions made by free energy minimization is limited by the quality of the energy parameters in the underlying free energy model. This paper proposes a new algorithm that computes base pairing pattern for RNA molecule by using stochastic context-free grammars (SCFGs). Complex internal structures in RNA and equilibrium concentrations of duplex structures are fully taken into account. This new algorithm is compared with dynamic programming benchmark mfold and algorithms (Tfold, and MaxExpect). The results showed that the proposed algorithm achieved better performance with respect to sensitivity and positive predictive value.

KEYWORDS: RNA Folding, RNA Secondary Structure, Computational Biology, Stochastic Context-Free Grammars (SCFGs), Sensitivity, Positive Predictive Value

INTRODUCTION

The hydrogen bonding and stacking interactions of the hydrophobic nucleobases are major contributors to the stable association of nucleotides within and between ucleic acid molecules. Hydrogen bonds are principally characterized by highly specific electrostatic interactions that stabilize the nucleic acid secondary structure. Watson–Crick hydrogen bonds between the bases of the nucleosides adenosine (A) and uridine (U) or thymidine, and guanosine (G) and cytidine (C), and a multitude of noncanonical hydrogen bonds play crucial roles in both the secondary and tertiary structures of nucleic acids and in their functions [1]. Accurate RNA secondary structure predictions help in understanding RNA's function, in identifying novel functional RNAs in genome sequences, and in recognizing evolutionarily related RNAs in other organisms. Most RNA secondary structure prediction algorithms are based on energy minimization [2]. RNA is a single strand of nucleotides composed of adenine (A), guanine (G), cytosine (C) and uracil (U) and it can fold back on itself to form its secondary structure with base pairs like A - U, G = C, and G ≡ U. However, an RNA sequence can fold to form several possible secondary structures. Determining the correct secondary structure is called the RNA secondary structure prediction problem [3]. The importance of Ribonucleic Acid (RNA) has increased in the recent years. It was found that RNA performs a central role within the living cells such as carrying genetic information (mRNA), interpreting the code (ribosomal RNA), and transferring genetic code (tRNA). It also plays many diverse roles in biology, include catalyzing chemical reactions [4],[5],[6], directing the site specific modification of RNA nucleotides, controlling gene expression, modulating protein expression and serving in protein localization [7],[8],[9]. On the other hand the importance of accurate predictions of secondary structures has increased due to the recent finding of functional non-coding RNAs whose functions are closely related to their secondary structures. Identifying the secondary structure of an RNA molecule

is the fundamental key to understand its biological function and predict its tertiary structure [10], [11], [12].The physical methods for RNA secondary structure prediction are time consuming and expensive, thus methods for computational prediction will be a proper alternative. Various algorithms have been used for RNA structure prediction including dynamic programming and metaheuristic algorithms [13], [14]. Computational methods for modeling RNA secondary structure provide useful initial models for solving the tertiary structure[15],[16]. The problem of computational prediction of secondary structure for a single RNA sequence dates back to the early 1970s. Zuker et al. [3] proposed the Dynamic Programming (DP) algorithm which called mfold. It is still a popular algorithm to find secondary structure of an RNA molecule. Moreover, it has become the benchmark for predicting the RNA secondary structure. mfold takes the primary RNA sequence as input, and uses a complex thermodynamic model to evaluate the free energy of the structures by seeking the pseudoknot-free secondary structure with the Minimum Free Energy (MFE). RNAFold from the ViennaRNA [17] package predicts the RNA secondary structure through energy minimization. It reads an RNA sequence as input and provides three kinds of algorithms to predict the structure: i) the MFE algorithm to find a single optimal structure; ii) the partition function algorithm to calculate the base pair probabilities in the thermodynamic ensemble; iii) the suboptimal folding algorithm to generate all suboptimal structures based on MFE. Currently, three different approaches study the RNA secondary structures. The first one is the single sequence approach which predicts the secondary structure by searching the minimum free energy (MFE) [15], [18]. The second is the comparative sequence analysis. The iterative process takes a sequence, apples accurate sequence alignments data and analyzes the structure that is common to all the sequences in the database. Most of the developed methods which based on the free energy minimization either apply Dynamic Programming (DP) or a metaheuristics on the domain. The third is Stochastic Context-Free Grammar (SCFG) which predicts the most possible RNA secondary structure using context-free grammar and a defined set of probabilities for each grammar rule. These algorithms form the base of using computer programs to predict RNA secondary structures [19]. Stochastic context-free grammars (SCFGs) have emerged as an alternative probabilistic methodology for modeling RNA structure [20]. These models specify formal grammar rules that induce a joint probability distribution over possible RNA structures and sequences. In particular, the parameters of SCFG models specify probability distributions over possible transformations that may be applied to a ‘‘nonterminal’’ symbol, and thus are subject to the standard mathematical constraints of probability distributions (i.e. parameters may not be negative, and certain sets of parameters must sum to one). Though these parameters do not have direct physical interpretations, they are easily learned from collections of RNA sequences annotated with known secondary structures, without the need for external laboratory experiments [21],[22]. A method to predict the conserved secondary structures from multiple aligned sequences was presented in [23]. This method first calculates the base pairing probability matrix for each sequence, which are subsequently averaged to yield the base pairing probability matrix of the alignment. The consensus secondary structure is obtained by maximizing the expected accuracy of the structure with respect to the base pairing probabilities. CONTRAfold [24], uses probabilistic parameters learned from a set of RNA secondary structures to predict base-pair probabilities and then predicts structures using the maximum expected accuracy approach. MaxExpect [25] predicts both the optimal structure (having highest expected pair accuracy) and suboptimal structures to serve as alternative hypotheses for the structure. It explores the use of maximum expected accuracy in single sequence secondary structure prediction, where thermodynamics are utilized to predict the underlying base-pair probabilities. Tfold [26] takes as input a RNA sequence for which the secondary structure is searched and a set of aligned homologous sequences. It combines criteria of stability, conservation and covariation in order to search for stems and pseudoknots (whatever their type). Stems are searched recursively, from the most to the least stable. It uses an algorithm called SSCA for selecting the most appropriate sequences from a large set of homologous sequences (taken from a database for example) to use for the prediction. This paper proposes a new folding algorithm that predicts the RNA

secondary Structure. Maximizing the number of base pairing and RNA stochastic context-free grammars (SCFGs) are taken into account during the program design. Also the rules that control RNA reliability are considered. Performance of the considered algorithm is compared with mfold, Tfold, and MaxExpect algorithms using standard sets of RNA test molecules. RNA secondary structure prediction using Stochastic Context Free Grammars is presented in Section 2. The proposed folding algorithm is introduced in Section 3. The experimental results are presented in Section 4. Conclusion is given in Section 5.

STOCHASTIC CONTEXT-FREE GRAMMARS AIDED SECONDARY STRUCTURE PREDICTION

One of the most promising techniques in bioinformatics is the analysis of stochastic grammars, since they allow the generation of sequence patterns in a natural way, besides having a broader range of action than other architectures [27], [28]. Stochastic grammars have its origins in formal grammars that were developed as a model to analyze natural languages. Grammars are useful tools to model character sequences, in a certain way are useful to model molecular biology sequences. Many bioinformatics problems can be reformulated in terms of formal languages, producing the corresponding grammar from the available data [29],[30].

Modeling Secondary Structure with SCFGs

In the RNA secondary structure prediction problem, we are given an input sequence x, and our goal is to predict the best structure y. For probabilistic parsing techniques, this requires a way to calculate the conditional probability $P(y / x)$ of the structure y given the sequence x.

Representation Stochastic Context -Free Grammars (SCFGs)

SCFGs provide a compact representation of a joint probability distribution over RNA sequences and their secondary structures. SCFG for secondary structure prediction defines (1) a set of transformation rules, (2) a probability distribution over the transformation rules applicable to each nonterminal symbol, and (3) a mapping from parses (derivations) to secondary structures. For example, consider the following simple unambiguous SCFG for a restricted class of RNA secondary structures [31],[32]:

- **Transformation Rules.**

  $$S \rightarrow aSu \mid uSa \mid cSg \mid gSc \mid gSu \mid uSg \mid aS \mid cS \mid gS \mid uS \mid \epsilon :$$

- **Rule Probabilities**. The probability of transforming a nonterminal S into aSu is $pS \rightarrow aSu$, and similarly for the other transformation rules.

- **Mapping from Parses to Structures**. The secondary structure y corresponding to a parse σ contains a base pairing between two letters if and only if the two letters were generated in the same step of the derivation for σ.

  For a sequence x = agucu with secondary structure y = ((.)), the unique parse s corresponding to y is

  $$S \rightarrow aSu \rightarrow agScu \rightarrow aguScu \rightarrow agucu.$$

  The SCFG models the joint probability of generating the parse s and the sequence x as

  $$P(x, \sigma) = P_{S \rightarrow aSu} \cdot P_{S \rightarrow gSc} \cdot P_{S \rightarrow uS} \cdot P_{S \rightarrow}$$

  It follows that

$$P(y|x) = \sum_{\sigma \epsilon y} P(\sigma|x) = \frac{\sum_{\sigma \epsilon y} P(x|\sigma)}{\sum_{\sigma' \epsilon \Omega(x)} P(x|\sigma')}$$

Where Ω(x) is the space of all possible parses of x.

RNA Secondary Structure

A secondary structure Y on a sequence x of length n is a set of base pairs (i, j), i <j, such that (i, j) ∈ Y implies that (xi, xj) is either a Watson-Crick (GC or AU) or a wobble (GU) base pair.

Every sequence position *i* takes part in at most one base pair, i.e., *Y* is a matching in the graph of "legal" base pairs that can be formed within sequence *x*.

- $(i, j) \in Y$ implies $|i - j| \geq 4$, i.e., hairpin loops have at least three unpaired positions inside their closing pair.
- If $(i, j) \in Y$ and $(k, l) \in Y$ with $i < k$, then either $i < j < k < l$ or $i < k < l < j$. This condition rules out knots and pseudoknots. Together with condition 1 it implies that *Y* is a circular matching [16, 25,33].

Secondary Structure Prediction Evaluation Method

Secondary structure prediction can be benchmarked using sensitivity and positive predictive value (PPV) for base-pair prediction. Sensitivity is the percentage of known pairs correctly predicted, and PPV is the percentage of predicted pairs that are in the known structure. These two statistics are calculated as:

$$Sensitivity = \frac{number\ of\ correctly\ predicted\ base\ pairs}{total\ number\ of\ known\ base\ pairs},$$

$$PPV = \frac{number\ of\ correctly\ predicted\ base\ pairs}{total\ number\ of\ predicted\ base\ pairs}$$

The sensitivity of free energy minimization has been benchmarked as high as 73% on a diverse database of RNA sequences with known structures of fewer than 700 nucleotides [33], [34], but the PPV of free energy minimization is only 66%. The lower PPV has two causes. First, there is a tendency to over predict base pairs because the formation of pairs lowers the free energy change. Second, occasionally the database of known structures does not annotate all experimentally determined base pairs, and predicting a correct pair that is not annotated lowers PPV [35].

THE PROPOSED FOLDING ALGORITHM

RNA structure is dominated by base pairs which induce highly predictable patterns of long distance pairwise residue complementarily in RNA primary sequence. A stochastic grammar specifies a probability for each production in the grammar and thus assigns a probability to each string (sequence) it derives. This paper introduces the folding algorithm which implements a SCFG for RNA secondary structure predication. The stability of the RNA secondary structure increases according to the number of GC versus AU and GU. Also the number of base pairs in a hairpin loop region has a high effective in RNA stability. The number of unpaired bases decreases the stability of the structure such as interior loops, hairpin loop and bulges. So increasing the predicted base pairs is taken as a parameter in the proposed algorithm. Given a single RNA sequence, first starting from position i, scanning the total RNA and finding all possible base pairs for nucleotide i. assigning a number for each nucleotide j that can base pair with nucleotide i wight(i, j). This number is calculated by knowing the maximum number of base pair that can be formed if this pairs are joined. Second, The base-pair probability of an i-j pair, Pbp(i, j), is calculated using position-specific SCFG probability. This proposed wants to search a

sequence database for similar sequences. A position-specific SCFG is constructed which has a set of 4 scores for each single-stranded position, 16 scores for each base pair, and appropriate extra states and state transitions that allow for insertions and deletions. The base-pair probabilities provide the information about base pairs shared by multiple secondary structures. The base pairs with higher base-pair probability are more likely to be correctly predicted than base pairs with lower base-pairing probability. Accordingly, The final probability of the base pair i,j can be calculated using

$$final\ probability\,(i,j) = \frac{[wight(i,\ j)]^{\alpha} + [Pbp(i,j)]^{\beta}}{\sum_{all\ possible\ j}[wight(i,\ j)]^{\alpha} + [Pbp(i,j)]^{\beta}}$$

α, and β are two variables that used to determine the percentage contribution of every parameter in the final probability. The proposed algorithm was run with different values for α, and β. The best result was found when α equal to 0.7 and β equal to 0.3. Base pairs in the proposed folding RNA algorithm must satisfy the following constraints:

1) For (i,j), it must be canonical base pairs;

2) Each base cannot share more than one base;

3) Pairing bases must be at least three bases apart $i-j > 3$

4) two base pairs must not cross, i.e.: i, j ∩ i' , j' = ∅ or for all (i, j), (i' , j') either i < i' < j < j' hold.

The pseudo code of the proposed folding algorithm is shown in figure 1

```
-Input RNA sequence.
-Calculate number of nucleotides in the given RNA sequence (L).
-Starting from 5' end.
-Set i = 1.
-For i = 1 to L
     - find all possible canonical base pairs pools.
     - if nucleotide (i- 1) connected with nucleotide (j)
          -  search in canonical base pairs of nucleotide i for nucleotide (j+1)
          -  if  nucleotide  (j+1) found
               -  connect nucleotide  i to nucleotide j+1
           end if
      else
          - calculate weight and base pair probability for every nucleotide in the pool
            then calculate final probability .
          - find the nucleotide that give maximum final probability
          - check RNA constraints. If they are satisfied then connect nucleotide i to
            the chosen nucleotide
       end if
end for
- Return final chosen secondary structure
```

Figure 1: The Pseudo Code of the Proposed Folding Algorithm

EXPERIMENTAL RESULTS

Secondary structures determined by comparative analysis were used as known structures. The database includes various types of RNA as used in previous benchmarks [23, 34]. Also the test sequences were taken from the comparative RNA website [35]. Maximization of expected final probability provides a best estimate for an RNA secondary structure. The base-pair probabilities and Wight value are calculated from actual information provided by the studying RNA. The proposed algorithm was tested on a diverse database of RNA sequences with known secondary structures. The optimal, folding structure is predicted and compared with the known structure in the database. The accuracy of prediction is reported as sensitivity and PPV. Table 1, summarizes the accuracies of the proposed algorithm, Tfold, MaxExpect, and free energy minimization for RNAs in the database. The proposed algorithm was tasted with different values for parameters α, and β. The best result with respect to sensitivity is reported in table 1, for α =0.7, and β = 0.3. The result shows that the contribution of weight in determining the final probability must be more than the

contribution of base pair probability. That is because weight calculation based on maximizing the number of base pairs which lead to increase RNA stability.

Table 1: The Accuracies of the Proposed Algorithm, Tfold, MaxExpect, and Free Energy Minimization

Type of RNA	Proposed Algorithm		Tfold		MaxExpect		Free Energy Minimization	
	Sensitivity %	PPV %	Sensitivity %	PPV %	Sensitivity %	PPV %	Sensitivity %	PPV %
5s rRNA	85±15	92±7.0	80±15.0	93±7.0	72.5 ± 26.4	65.3±23.6	72.9±26.6	63.9±23.8

The table result shows that the proposed algorithm achieved best results with respect to Tfold, MaxExpect and mfold algorithms. That is because of joining the SCFG probability with max base pair weight. Each parameter of them contributed in finding the best base pairing. Figure 2 and 3, show the secondary structure predicted by the proposed algorithm compared with the known secondary structure. The blue lines represent the base pairs for both the known and predicted structure. The red lines represent the predicted base pair which is not found in the known structure. The green lines indicate the base pairs in the known structures which have not been predicted. It is noted that the proposed algorithm in figure 2, could find 30 base pairs out of 31, which is 96.8% of the known base pairs. While the result regard to mfold was 94.1%. In figure 3, the proposed algorithm could find all base pairs correctly. While the result regard to mfold was 87%.

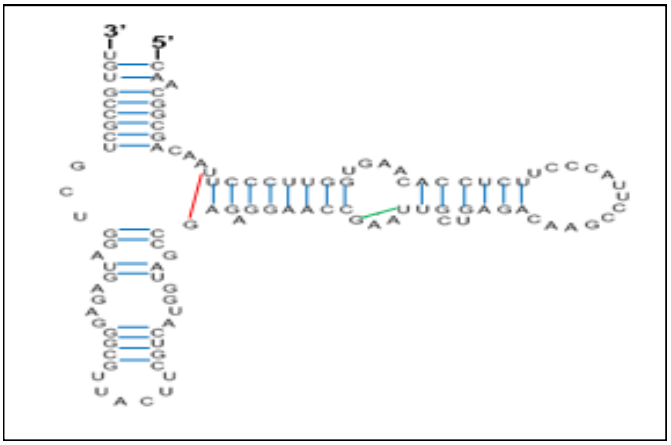

Figure 2: The Secondary Structures of 5S Ribosomal RNA (X14441)

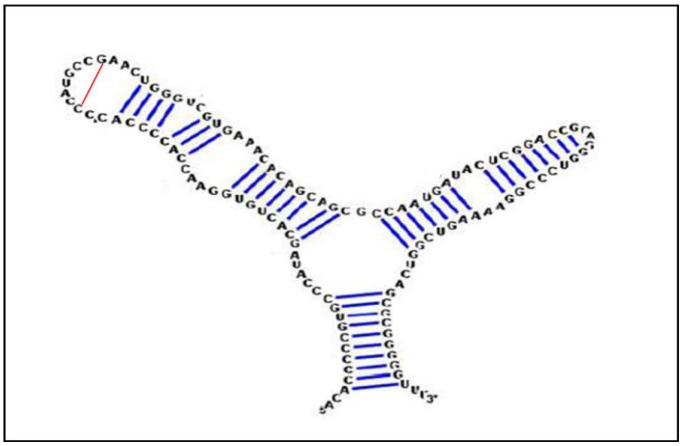

Figure 3: The Secondary Structures of 5S Ribosomal RNA (X67579)

CONCLUSIONS

Because the chemical and biological properties of many RNAs are determined by their conformations, an RNA equivalent exists to the protein folding problem. The RNA folding problem has both a practical side and an academic side. If the practical problem were solved, it would be possible to predict the conformations of RNAs of known sequence. In this paper, a new proposed algorithm is presented to find the RNA secondary structure. It is depended on joining the SCFG probability with max base pair weight. The proposed algorithm provides better performance in predicting the RNA secondary structure. Comparisons to mfold, Tfold, and MaxExpect algorithms have been performed. The proposed algorithm is found to be achieved better performance than the other algorithms. That is because it tries to increase the predicted RNA stability by increasing the total base pairs. The comparisons between those algorithms were in terms of sensitivity and positive predictive value.

REFERENCES

1. M. Zuker & D. Sanko. (1984) RNA secondary structures and their prediction. Mathematical Bioscience, Vol. 46, pp. 591-621.
2. Nussinov R. & Jacobson, A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Natl. Acad. Sci, 77: 6309–6313.
3. Zuker, M. & Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., 9: 133–148.
4. J. A. Doudna & T. R. Cech. (2002) The chemical repertoire of natural ribozymes. Nature, vol. 418, no. 6894, pp. 222–228.
5. J. L. Hansen & et al. (2002) Structural insights into peptide bond formation, in Proceedings of the National Academy of Sciences, vol. 99, pp. 11 670–11 675.
6. Nissen, P. & et al. (2000) the structural basis of ribosomal activity in peptide bond synthesis. Science, 289: 920–930.
7. D. H. Mathews. (2006) Revolutions in rna secondary structure prediction. Journal of Molecular Biology, vol. 359, p. 526532.
8. G. Meister & T. Tuschl. (2004) Mechanisms of gene silencing by double stranded rna. Nature, vol. 431, pp. 343–349.
9. Walter, P. & Blobel, G. (1982) Signal recognition particle contains a 5S RNA essential for protein translocation across the endoplasmic reticulum. Nature 982, 299: 691–698.
10. H. Tsang & K. Wiese. (2007) Sarna-predict: A study of rna secondary structure prediction using different annealing schedules. In Proceedings of the IEEE Symposium on Computational Intelligence in Bio-M. informatics and Computational Biology, pp. 239–246.
11. Neethling and A. Engelbrecht. (2006) Determining rna secondary structure using set-based particle swarm optimization. In IEEE Congress on Evolutionary Computation (CEC2006), pp. 1670–1677.
12. Do,C., Foo,C. & Batzoglou,S. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics, 24, 68–76.

13. Lagos-Quintana & et al. (2001) Identification of novel genes coding for small expressed RNAs. Science 294: 853–858.

14. Xia, T. & et al. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. Biochemistry, 37: 14719–14735.

15. Douglas H. Turner1 & David H. Mathews. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Research, Vol. 38

16. Song Cao & Shi-Jie Chen. (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA Society, 11: 1884-1897.

17. Ivo L. Hofacker. (2003) Vienna RNA secondary structure server. Nucleic Acids Research, Vol. 31, No. 13.

18. Layton, D.M. & Bundschuh, R. (2005) A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. Nucleic Acids Res, 33: 519–524.

19. Zsuzsanna Sükösd & et.al. (2011) Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. BMC Bioinformatics, 12:103.

20. R. Nussinov. & et.al. (1978) Algorithms for loop matching's. SIAM Journal on Applied Mathematics, vol. 35, no. 1, pp. 68–82.

21. Eddy SR. (2001) Non-Coding RNA Genes and the Modern RNA World. Nat. Rev. Genet., 2:919–929.

22. Harmanci,A.O. & et.al. (2008) PARTS: probabilistic alignment for RNA joint secondary structure Prediction. Nucleic Acids Res. 36, 2406–2417.

23. Eddy SR. & Durbin R. (1994) RNA Sequence Analysis Using Covariance Models. Nucl. Acids Res., 22:2079–2088.

24. Chuong B. D. & et. al. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. BIOINFORMATICS, Vol. 22 no. 14, pages e90–e98.

25. Zhi John Lu & et.al. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA Society, Vol. 15, No. 10.

26. Ste´ fan Engelen & Fariza Tahi. (2010) Tfold: efficient in silico prediction of non-coding RNA secondary structures. Nucleic Acids Research, Vol. 38, No. 7.

27. Tinoco Jr. & et.al. (1971) Estimation of secondary structure in ribonucleic acids. Nature, 230: 362–367.

28. Justin T. Low & Kevin M. Weeks. (2010) SHAPE-Directed RNA Secondary Structure Prediction. Elsevier, 52(2): 150–158.

29. Lisa Yu & et. al. (2006) Study of RNA Secondary Structure Prediction Algorithms. Master thesis, San Jose State University.

30. YE DING. (2006) Statistical and Bayesian approaches to RNA secondary structure prediction. RNA Society, Vol. 12, No. 3.

31. Abdulqader M. Mohsen & et.al. (2009) Predicting the minimum free energy RNA Secondary Structures using Harmony Search Algorithm. World Academy of Science, Engineering and Technology 56.

32. Micheal P. S. Brown (2000) Small Subunit Ribosomal RNA Modeling using Stochastic Context Free Grammars. American Association for Artificial Intelligence.

33. Mathews DH & Turner DH. (2006) Prediction of RNA secondary structure by free energy minimization. Curr. Opin. Struct. Biol., 16: 270–278.

34. Mathews & et.al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci, 101: 7287–7292.

35. J. J. Cannone, S. Subramanian & et. al. (2002). The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. BMC Bioinformatics, vol. 3, no. 35, pp. 169–172.